ETS

Listening. Learning. Leading.®

Research Spotlight

Issue 1 • February 2008

ETS Research Spotlight
Volume 1

Copies can be downloaded from:
www.ets.org/research

February 2008
Research & Development Division
Educational Testing Service

To comment on any of the articles in this issue, write to ETS Research Communications via e-mail at: R&DWeb@ets.org

Table of Contents

Foreword

During the course of 2007, ETS researchers and scientists published 33 articles in refereed journals, 57 book chapters, edited or co-edited seven books, and published 44 reports in the ETS Research Report Series. In addition, our researchers gave more than 400 presentations at conferences around the world.

All of this activity speaks to the energy and professionalism of our staff. It also makes it virtually impossible to capture the full range of the research ETS undertakes each year as we work to fulfill the organization's mission "to help address quality and equity in education." The Research & Development Division, however, has initiated the Research Spotlight Series in order to provide a small taste of our work to our colleagues in the field of educational measurement. For a wider view of our various research activities, I encourage you to visit www.ets.org/research and the ETS ReSEARCHER (http://search.ets.org/custres/), where you can search the entire ETS Research Report Series. The data base has the abstracts for some 3,000 ETS research reports dating back to 1948. Many of the newer reports are available at no cost as PDFs.

If you have questions about any of the articles or our research portfolio, please contact us. You can send your inquiry via e-mail to R&DWeb@ets.org.

Ida Lawrence
Senior Vice President
ETS Research & Development

Acknowledgments

The inaugural issue of the ETS Research Spotlight was reviewed by Daniel Eignor, Principal Research Scientist at ETS. The report was edited by William Monaghan, Manager of Proposal Development within the Research & Development Division at ETS. He also provided the desktop publishing. Data used in "Reading Aloud as an Accommodation for a Test of Reading Comprehension" was derived in part from the National Accessible Reading Assessments Projects (NARAP, www.narap.info), which is a collaborative effort funded by the U.S. Department of Education. Errors of fact or interpretation are those of the authors. All pictures by William Monaghan.

Choice of Anchor Test in Equating

Sandip Sinharay and Paul Holland

Editor's note: *ETS continually engages in studies to improve the quality and reduce the cost of the testing process. To ensure that scores on two or more forms of a test can be used interchangeably, people in the testing field turn to equating, which can be a labor-intensive task. A common equating practice is to use an anchor test (a miniature version of the tests being equated) for this purpose. In the following article, ETS researchers describe their study that suggests that contrary to common knowledge, the items in an anchor test do not need to reflect the full range of difficulty of the items in the tests being linked. The implication is that it will take less effort for developers of certain tests to create effective anchor tests.*

The Non-Equivalent groups with Anchor Test (NEAT) design is one of the most flexible tools available for equating tests (e.g., Kolen & Brennan, 2004). The NEAT design deals with two non-equivalent groups of examinees and an anchor test. The design table for a NEAT design is shown in Table 1.

Table 1. **The NEAT Design**

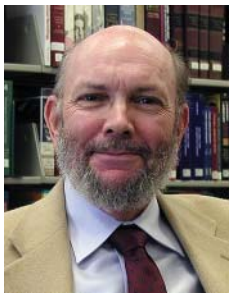
		Tests		
		X	A	Y
Population	P	×	×	
	Q		×	×

The test X corresponds to the new form given to a sample from population P and the test Y corresponds to the old form given to a sample from population Q. The anchor test A is given to both P and Q. The choice of anchor test is crucial to the quality of equating with the NEAT design.

It is a widely held belief that an anchor test should be a miniature version (i.e., a minitest) of the tests being equated. It is recommended that an anchor test be proportionally representative or a mirror of the tests to be equated in both content and statistical characteristics (Kolen & Brennan, 2004, p. 19). Currently, most operational testing programs that use the NEAT design employ a minitest as the anchor. To ensure statistical representativeness, the usual practice is to set the mean and spread of the item difficulties of the anchor test so that they are roughly equal to those of the tests being equated.

Sandip Sinharay

The requirement that the anchor test be representative of the tests to be equated with respect to content is justified from the perspective of content validity and has been shown to be important by researchers like Klein and Jarjoura (1985). Peterson, Marco, and Stewart (1982) demonstrated the importance of having the mean difficulty of the anchor tests be close to that of the tests to be equated. We also acknowledge the importance of these two aspects of an anchor test. However, the literature does not offer any proof of the superiority of an anchor test for which the spread of the item difficulties is representative of the tests to be equated. Furthermore, a minitest has to include very difficult or very easy items to ensure adequate spread of item difficulties, which can be problematic as such items are usually scarce (one reason being that such items often have poor statistical properties like low discrimination and are thrown out of the item pool). An anchor test that relaxes the requirement on the spread of the item difficulties might be more operationally convenient, especially for testing programs using external anchor tests.



Paul Holland

Motivated by the above, this paper focuses on anchor tests that

- are content representative
- have the same mean difficulty as the tests to be equated
- have spread of item difficulties less than that of the tests to be equated

Operationally, such an anchor test can be constructed exactly in the same manner as the minitests are constructed except for the requirement that it mimic the spread of the item difficulties of the tests to be equated. Because items with moderate difficulty values are usually more abundant, an operationally convenient strategy to construct such an anchor test will most often be to include several moderate-difficulty items in the anchor test.

To demonstrate the adequate performance of anchor tests with spread of item difficulties less than that of the minitest, Sinharay and Holland (2006) defined a "miditest" as an anchor test with a very small spread of item difficulties and a "semi-miditest" as one with a spread of item difficulty that lies between those of the miditest and the minitest. These anchor tests, especially the semi-miditest, will often be easier to construct operationally than minitests because there is no need to include very difficult or very easy items in them. Sinharay and Holland cited several works that suggest that the miditest will be satisfactory with respect to psychometric properties like reliability and validity. This paper performs the next step, that of comparing the equating-performance of minitests versus that of miditests and semi-miditests through a pseudo-data example.

A Pseudo-data Example

It is not easy to compare a minitest versus a semi-miditest in operational setting as almost all operational anchor tests are constructed to be minitests. However, a study by von Davier, Holland, and Livingston (2005) allowed us perform the comparison. The study considered a 120-item test given to two different examinee samples P and Q of sizes 6168 and 4237 respectively. The sample Q has a higher average score, by about a quarter in standard-deviation-of-raw-score unit. Two 44-item tests X and Y, as well as anchor tests (that were constructed to be minitests) of lengths 16, 20, and 24 were constructed by partitioning the 120-item test. The 20-item anchor was a subset of the 24-item anchor and the 16-item anchor was a subset of the 20-item anchor. The test X was designed to be much easier (the difference being about 128% in standard-deviation-of-raw-score unit) than the test Y.

Of the total 120 items in the test, items 1-30 are on language arts, 31-60 are on mathematics, 61-90 are on social studies, and 91-120 are on science. As the minitest, we take the 16-item anchor test of von Davier et al. (2005). There were not enough middle-difficulty items to choose a miditest. The semi-miditest we chose was a subset of the 24-item anchor test of von Davier et al. We ranked the six items within each of the four content areas in the 24-item anchor test according to their difficulty (proportion correct). The four items ranked 2nd to 5th within each content area were included in the 16-item semi-miditest. Nine items belonged to both the minitest and semi-miditest. We refer to this example as a "pseudo-data" example rather than a "real data" example because the tests to be equated and the anchor tests we consider were not operational, but artificially constructed from real data.

Note that by construction, the semi-miditest, like the minitest, is content representative. Also, the semi-miditest has roughly the same average difficulty as the minitest; the average difficulties of the minitest and the semi-miditest are 0.68 and 0.69 respectively in P, and 0.72 and 0.73 respectively in Q. However, the spread of the item difficulties of the semi-miditest is less than that of the minitest. For example, the standard deviation of the item difficulties of the minitest and the semi-miditest are 0.13 and 0.09 respectively in P, and 0.12 and 0.08 in Q (the standard deviation of the item difficulties for X in P is 0.12 while that for Y in Q is 0.17).

The first two rows of Table 2 show the anchor-to-test correlation coefficients.

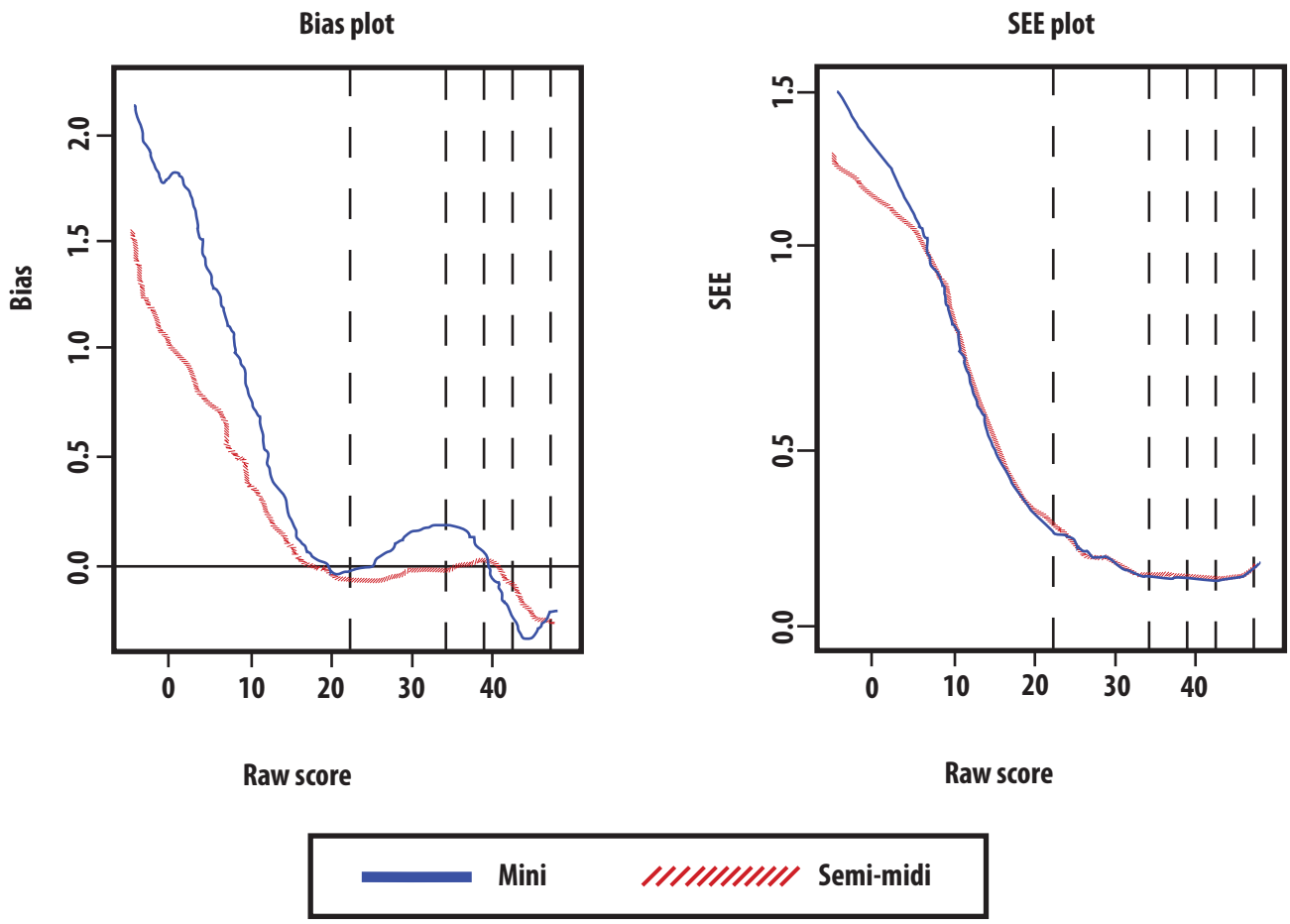
Table 2. **Findings from the Long Basic Skills Test**

	Minitest	Semi-miditest
Correlation for X and A in P	0.75	0.73
Correlation for Y and A in Q	0.75	0.68
Weighted average of absolute bias	0.18	0.08

We computed the equating functions for equating X to Y using the chain equipercentile method, using presmoothing and linear interpolation, for the minitest and the semi-miditest, by pretending that scores on X were not observed in Q and scores on Y were not observed in P (i.e., treating the scores on X in Q and on Y in P as missing). We also computed the criterion ("true") equating function for equating X to Y by employing a single-group equipercentile equating with linear interpolation on all the data from the combined sample of P and Q.

Figure 1 shows a plot of the bias and standard error in equating

Figure 1. **Plot of the bias and SEE for equating X to Y for the semi-miditest and minitest**



(SEE) for equating X to Y for the semi-miditest and minitest. The bias here is defined as the difference between the equating function and the above-mentioned criterion equating function. Each panel of the figure also shows, using vertical lines, the five quantiles, for p = 0.025, 0.25, 0.50, 0.75, 0.975, of the scores on X in the combined sample including P and Q.

Table 2 also shows weighted averages of absolute equating bias, the weight at any score point being proportional to the corresponding frequency in the combined sample. There is little difference between the minitest and the semi-miditest with respect to equating bias and SEE, especially in the region where most of the observations lie. The SEEs for the two anchor tests are very close while the semi-midi test has somewhat less bias. Contrary to the intuitive expectation of some of our colleagues, Figure 1 shows that at the extreme scores the SEE obtained by using a minitest is not smaller than the SEE obtained by using a semi-miditest.

Thus, the pseudo-data example, even with its limitations, such as short test and anchor test lengths and large difference between the tests to be equated, provides us with some evidence that a semi-miditest does not perform any worse than a minitest in operational equating.

Detailed simulation studies performed by Sinharay and Holland (2007), using data generated both from a unidimensional IRT model and a multidimensional IRT model, demonstrate that the equating-performance of miditests and semi-miditests is at least as good as that of minitests.

Conclusions

Our results suggest that the requirement of an anchor test to have the same spread of item difficulty as the tests to be equated may be too restrictive and need not be optimal. The design of anchor tests can be made more flexible, rather than employing the use of minitests, without losing any important statistical features in the equating process. Our recommendation then is to enforce a restriction on the spread of item difficulties of an anchor test only when it leads to operational convenience. For example, for tests using internal anchors, using a minitest (i.e., restricting the spread to be the same as that of the tests to be equated) may be more convenient because the scarce extreme difficulty items can be used in the anchor test and hence in both of the tests to be equated. For external anchors, our recommendation is to worry about content, average difficulty, and any other requirement, but not about spread of difficulty.

References

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. *Journal of Educational Measurement, 22,* 197-206.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating method. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71-135), New York: Academic Press.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an anchor test* (ETS Research Report No. RR-06-04). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Livingston, S. A. (2005). *An evaluation of the kernel equating method: A special study with pseudo-tests from real test data.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QU.

Validating the Use of TOEFL® iBT Speaking Section Scores to Screen and Set Standards for International Teaching Assistants

Xiaoming Xi

Editor's note: *For ETS, the appropriate use of test scores is paramount. The ETS Standards for Quality and Fairness (Educational Testing Service, 2002) state that there should be "evidence that indicates that the product or service is likely to meet the intended purpose for the intended population (p. 11)." So in being responsive to the needs of the educational marketplace, ETS strives to be responsible by ensuring its products and services fit their intended purpose(s) and provide truly meaningful information to help the users make score-based decisions. The study below is such an example, in which a researcher investigates the suitability of using the test scores for a secondary purpose and provides guidance in how to use the scores for decision-making in the local contexts.*

The Test of English as a Foreign Language™ (TOEFL®) has undergone major revisions, including the introduction of speaking as a mandatory section on the TOEFL Internet-based test (iBT). The TOEFL iBT Speaking Section has been designed to measure a candidate's ability to communicate orally in English in an academic environment. Although it is used primarily to inform admission decisions regarding international applicants at English medium universities, it may also be useful as an initial screening measure for international teaching assistants (ITAs).

As reviewed in Plakans & Abraham (1990), three major types of tests have been used to test the oral skills of ITAs: the Test of Spoken English™ (TSE®) or SPEAK® developed by ETS, oral interviews, and teaching simulation tests. These tests have served complementary functions in ITA testing. Traditionally, the TSE, administered in test centers around the world, has served the purpose of pre-arrival screening. However, while the TSE uses speaking tasks that are contextualized in more general communicative settings, the TOEFL iBT Speaking Section has been designed specifically to measure oral communication skills for academic purposes. Thus, it may be a more appropriate measure for ITA screenings than the TSE, given its focus on academic contexts. In addition, the TSE has mostly been phased out with the launch of the TOEFL iBT test world wide, and a new pre-arrival screening test is needed.

Locally administered SPEAK exams, which use retired TSE forms, have been widely used as an on-site ITA screener, alone or along with locally developed teaching simulation tests. Although the TOEFL iBT test has been launched in the majority of locations worldwide, the SPEAK test can still be used for on-campus initial screening. It should be noted, however, that ETS no longer supports or carries this product. Nevertheless, for incoming international students who submit their TOEFL iBT scores with their applications (including their Speaking Section scores), the TOEFL iBT Speaking Section scores could potentially be used for pre-admission screening. Such an approach would aid in identifying candidates who are ready to teach as well as help determine who needs to be tested using the local test before and/or after they have arrived.

Xiaoming Xi

The goals of this study are to provide criterion-related validity evidence for ITA screening decisions based on the TOEFL iBT Speaking Section scores and to evaluate the adequacy of using the scores for TA assignments. First, this study investigates the relationships between scores on TOEFL iBT Speaking Section and scores on criterion measures, intending to establish some association between them. Then, it illustrates how cut scores for TA assignments can be determined based on students' performances on the TOEFL iBT Speaking Section and on the criterion measures.

In this study, two types of criterion measures for the TOEFL iBT Speaking Section were used: locally developed teaching simulation tests used to select ITAs and ITA course instructors' recommendations of TA assignments. Institutions which adopt fairly established procedures to select ITAs were selected. In

particular, they use various performance-based tests that attempt to simulate language use in real instructional settings. This type of teaching simulation test was considered to be more authentic in resembling the real-world language use tasks and in engaging the underlying oral skills required in instructional settings, in comparison to a tape-mediated general speaking proficiency test and oral interview (Hoekje & Linnell, 1994). At these participating institutions, various studies have been conducted to support the validity of their tests for ITA screenings or procedures have been established to check the effectiveness of the ITA test for ITA assignments. Whenever feasible, the reliability of the scores on a local ITA test was estimated in this study and then the observed correlation between the local ITA test scores and the TOEFL iBT Speaking Section scores was corrected for score unreliability to reveal the "true" relationships between them. Otherwise, measurement errors associated with scores on both the TOEFL iBT Speaking Section and the local ITA test may disguise the true relationships between them.

The most important focus of this paper is to illustrate the process of setting cut scores for ITA screenings. This involves both methodological considerations and value judgments. On the methodological side, it demonstrates how the overall effectiveness of TOEFL iBT Speaking Section scores in classifying TA assignments can be established by using binary or ordinal logistic regression (Agresti, 2002; Hosmer & Lemeshow, 2000). It also discusses two types of errors that may occur when using TOEFL iBT Speaking Section scores for classifying students for teaching assignments, taking into account their trade-offs, which reflect value judgments, in order to establish an appropriate standard in ITA screening.

Trade-off of Different Classification Errors in Using TOEFL iBT Speaking Section Scores for ITA Assignments

When TOEFL iBT Speaking Section scores are used to classify students for TA assignments, two types of classification errors are likely to occur: false positives and false negatives. In this context, false positives occur when those who are not qualified TAs based on their local ITA test scores are predicted to be qualified by their TOEFL iBT Speaking Section scores. In contrast, false negatives occur when candidates who are qualified TAs are predicted to be unqualified by their TOEFL iBT Speaking Section scores. Since ITA programs are gatekeepers for quality undergraduate education, false positives may have more serious impact, since having unqualified ITAs in classrooms may compromise the quality of undergraduate education and infringe on the interests of undergraduate students who pay high tuitions and fees.

The other factor to consider in setting cut scores is to what extent a specific type of error could be rectified. This study examines the use of TOEFL iBT Speaking Section scores as an initial screening measure to help identify qualified TAs. If an unqualified TA were classified as qualified by his/her TOEFL iBT Speaking Section score (a false positive), there would be no way to rectify this error. However, if an otherwise qualified ITA were predicted as unqualified (a false negative), he/she would still have a chance to be tested using the local ITA test once they arrived. The impact would be that his/her TA employment may be delayed until he/she passes the local test. After considering the potential impact of the two types of errors and how rectifiable they are, it was decided that it is more important to minimize false positives at the expense of false negatives.

The Study

Four universities participated in this study: University of California, Los Angeles (UCLA), University of North Carolina, Charlotte (UNCC), Drexel University (Drexel), and University of Florida at Gainesville (UF). At all these universities, an in-house ITA screening test has been used alone or in conjunction with the SPEAK test to screen ITAs. At each institution, students who signed up for their local ITA tests were invited to take the TOEFL iBT Speaking Section as well. Table 1 summarizes the data collected at the four institutions.

Table 1. **Data Collected at Each Participating School**

	TOEFL Speaking Section	In-house ITA test	SPEAK	Instructor recommendations
UCLA	×	×		
UNCC	×	×		
Drexel	×	×	×	×
UF	×	×	×	

The next section uses UCLA as an example to demonstrate the relationship between the local ITA test scores and TOEFL iBT Speaking Section scores and the process of setting cut scores on TOEFL iBT Speaking Section for ITA selection.

Illustrative example – UCLA

Local ITA Assessments and Requirements for ITAs

The Test of Oral Proficiency (TOP) has recently replaced SPEAK at UCLA for screening ITAs. It is a locally developed test that consists of three tasks: A self-introduction (not scored), a short-presentation on some typical classroom materials provided, and a prepared presentation about a basic topic in the examinee's own field.

The short presentation and the prepared presentation tasks are each double scored in an analytic fashion on Pronunciation, Vocabulary/Grammar, Rhetorical Organization, and Question Handling. The composite TOP score is derived by summing the four scores, with a 1.5 weight assigned to pronunciation. Then it is scaled to a range of 0 to10. A score of 7.1 or higher is necessary for a "clear pass" which will allow a student to work as a TA. A score of 6.4 to 7.0 is considered a "provisional pass", and students receiving scores in this range are required to take an ITA oral communications course prior to or during their first quarter of TA work. A score lower than 6.4 is not high enough to qualify for TA work.

Participants and Procedure

Eighty-four international graduate students who were roughly representative of the TOP examinee population at UCLA took both the TOP and TOEFL iBT Speaking Section. Forty-two (50.0%) of them were classified as clear passes, 15 (17.9%) as provisional passes and 27 (32.1%) as non-passes based on their TOP scores.

Relationships between TOEFL iBT Speaking Section Scores and TOP scores

Table 2 demonstrates that the correlations among TOEFL iBT Speaking Section scores and TOP composite and analytic scores were moderately high. After correcting for score unreliability, the correlation between the TOEFL iBT Speaking Section and TOP composite scores was .84. The disattenuated correlations, which were observed correlations corrected for score unreliability, also show that the TOEFL iBT Speaking Section scores had strong correlations with the TOP analytic scores, showing the strongest relationship with the TOP Grammar & Vocabulary scores (.86).

Overall Effectiveness of Using TOEFL iBT Speaking Section Scores for ITA Screening

Sixty-five cases, randomly selected from the whole sample, were used in model building and the remaining 19 cases were used in testing the classification accuracy. An ordinal regression model with a logit link satisfied the assumption of parallel regression lines and also provided good classification results. The results show that the TOEFL iBT Speaking Section scores were a significant predictor of the TA assignment outcomes.

The classification accuracy further demonstrates how the

Table 2. **Observed and Disattenuated Correlations between the TOEFL iBT Speaking Section Scores and TOP Scores (N=84)**

	TOP	TOP Pronunciation	TOP Vocabulary & Grammar	TOP Organization	TOP Question Handling
TOEFL iBT Speaking Section	.78	.75	.75	.68	.69
	.84	**.81**	**.86**	**.80**	**.82**

Note: The disattenuated correlations are in bold face.

Table 3. **True versus Predicted Outcome Categories at UCLA**

True TA assignment outcome	Predicted TA assignment outcome			Percentage correct
	Non-passes	Provisional passes	Clear passes	
Non-passes	18	1	3	81.8%
Provisional passes	5	3	2	30.0%
Clear passes	1	0	32	97.0%
			Overall percentage	81.5%

TOEFL iBT Speaking Section scores performed in classifying students into one of the three outcomes. In Table 3, cases on the diagonal were correctly classified and the off-diagonal ones represent incorrectly predicted cases. The model did a superb job of correctly classifying the clear passes (97.0%), fairly well with the non-passes (81.8%), but not as well with the provisional passes (30.0%). This may be due to the fact that the model employed many fewer cases in the provisional pass category. Further, these provisional pass students were borderline students and may be more difficult to classify accurately.

Setting the Cut Scores

In the ROC curve for provisional passes (Figure 1), the area under the curve was very high (.91), indicating that the probability of the TOEFL iBT Speaking Section score of a marginal or clear pass student exceeding that of a non-pass student was 91%. That is to say, if we randomly select a clear pass student and a non-pass student, 91% of the time, the TOEFL iBT Speaking Section score of the former will be higher than that of the latter. Table 4 contrasts the true positive and false positive rates for different TOEFL iBT Speaking Section score points for provisional passes. When the cut score is set at 24, no false positives will occur, but the true positive rate will stand at 53.5%. In other words, the model has to misclassify 46.5% of the marginal or clear passes as non-passes to correctly classify all non-passes. If 23 is chosen as the cut score, approximately five out of 100 non-passes may be classified as provisional passes. However, 11.6% (65.1% - 53.5%) more provisional passes will be correctly classified. This would reduce the number of students to be tested locally using the TOP but increase the number of students in ITA training classes. The slightly lower cut score (23) might be justified for two reasons: 1) Many science departments who hire the most ITAs are in dire need of TAs and a larger pool of eligible ITAs would help meet this need; 2) ITA course instructors can offer extra help in class to rectify the situation where non-passes are assigned TA work with concurrent English coursework on oral communication skills.

Using a table similar to Table 4, but for clear passes, 27 was estimated as the optimal cut score for identifying clear passes.

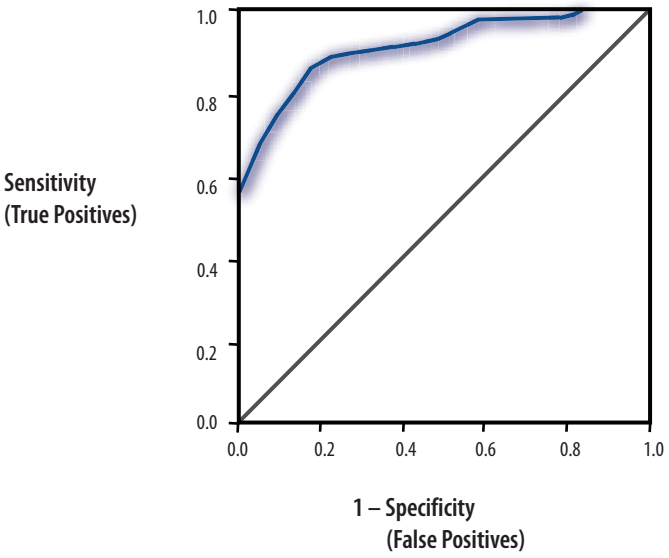


Figure 1. **The ROC Curve for Provisional Passes with TOEFL iBT Speaking Section Scores as the Predictor**

Table 4 . **True Positive Versus False Positive Rates at Different TOEFL iBT Speaking Section Cut Points for Provisional Passes**

Positive if Greater Than or Equal To	True Positive	False Positive
18.50	.884	.227
19.50	.860	.182
21.00	.791	.136
22.50	**.651**	**.045**
23.50	**.535**	**.000**
25.00	.395	.000
26.50	.279	.000

Note1: Not all possible cut points are displayed.

Note 2: The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All other cutoff values are the averages of two consecutive ordered observed test values. An integer cutoff value such as 21 is possible when the two consecutive test scores in the sample are 20 and 22. The cutoff values are rounded off to integers in the discussion of cut scores in the text because integer scaled scores are reported for the TOEFL iBT Speaking Section.

Cross-validation of the Classification Accuracy

The cut scores derived from the training sample were validated using the independent sample. As shown in Tables 5 and 6 using 23 or 24 as the cut score for provisional passes and 27 for clear passes, the classification accuracy with the independent sample was fairly similar: all the non-passes were correctly predicted;

Table 5. **Classification Rate on an Independent Sample with 27 on the TOEFL iBT Speaking Section for Clear Passes and 24 for Provisional Passes**

True TA assignment outcome	Predicted TA assignment outcome			Percentage correct
	Non-passes	Provisional passes	Clear passes	
Non-passes	5	0	0	100.0%
Provisional passes	4	0	1	0.0%
Clear passes	4	3	2	22.2%
			Overall percentage	36.8%

Table 6. **Classification Rate on an Independent Sample with 27 on the TOEFL iBT Speaking Section for Clear Passes and 23 for Provisional Passes**

True TA assignment outcome	Predicted TA assignment outcome			Percentage correct
	Non-passes	Provisional passes	Clear passes	
Non-passes	5	0	0	100.0%
Provisional passes	4	0	1	0.0%
Clear passes	3	4	2	22.2%
			Overall percentage	36.8%

only one of the provisional passes was incorrectly classified as a clear pass. However, some students were incorrectly classified into the lower categories. This is acceptable given that the false non-passes could be tested again using the local test once they arrive and those false provisional passes can get out of the ITA coursework at the recommendation of the instructors.

The false positive case causes some concern; however, UCLA allows provisional pass students to teach with concurrent enrollment in an English oral communication class. Therefore, if a mechanism is established for ITAs to receive some language support if it is found necessary after they start to teach, it should be reasonable to keep the cut score of 23 for provisional passes.

Summary of the Results from Four Institutions

Association between TOEFL iBT Speaking Section and Local ITA Test Scores

This study investigated the criterion-related validity of the TOEFL iBT Speaking Section scores for screening ITAs by examining its relationships with the local ITA test scores. The findings support the use of the TOEFL iBT Speaking Section

Table 7. **Correlations between the TOEFL iBT Speaking Section and the Local ITA Test Scores**

	UCLA TOP test	UNCC non-content presentation test	Drexel DIP test	UNCC content-based presentation test	UF Teach Evaluation
TOEFL iBT Speaking Section	.78	.78	.70	.53[2]	.44
	.84[1]	**.93**	Not available	**.58**	**.72**

Note 1: The disattenuated correlations are in bold face.

Note 2: This correlation was based on a very small sample (N = 23) and should be interpreted with caution.

Table 8. **Tasks and scoring rubrics of the local ITA tests**

	SPEAK	UCLA TOP test	UNCC Presentation tests	Drexel DIP test	UF Teach Evaluation
Tasks	Semi-direct test on topics of general or intellectual interest	Simulated teaching test (content and non-content combined)	Simulated teaching test (separate content and non-content based tests)	Simulated teaching test (content-based)	Real classroom teaching sessions
Rubrics	Linguistic	Linguistic	Linguistic Teaching	Linguistic Teacher presence & nonverbal communication	Linguistic Lecturing Cultural/Teaching

scores for ITA screening because the TOEFL iBT Speaking Section scores were reasonably correlated with scores on the local ITA screening measures.

As shown in Table 7, the TOEFL iBT Speaking Section scores had the strongest relationship with the UCLA TOP test scores and the UNCC Non-content Based Presentation test, less strong relationships with the Drexel DIP test scores and the UNCC Content-based Presentation test, and the weakest relationship with the UF Teach Evaluation scores. However, due to unavailability of data in some cases, some disattenuated correlations could not be estimated (e.g., Drexel). In other cases, the disattenuated correlations were underestimated as a result of the reliability of the local ITA tests being overestimated. Due to the particular assessment designs, such as single ratings of tasks or using a single task in an assessment, it was not possible to obtain appropriate reliability estimates that would take account of all potential sources of error (e.g., UNCC, Drexel and UF). In yet other cases, the restricted range of scores rendered the observed correlations lower than they would be if the whole range of possible scores were used (e.g., UF). Therefore, the disattenuated correlations provided only a partial picture of the "true" strengths of the relationships among these measures.

The strengths of the relationships were certainly affected by the extent to which the local ITA tests engaged and evaluated non-language abilities. As is evident in Table 8, the criterion measures used in this study certainly represent a continuum of less to more authentic tests. SPEAK can be placed on the left end of the continuum, since it uses tasks that are the least authentic in eliciting speech characteristic of language use in academic settings. The UCLA TOP test, the UNCC Presentation tests and the Drexel DIP test represent fairly authentic performance-based assessments that simulate the communication typical TA duties involve. On the right end of the continuum is the UF Teach Evaluation, which is an evaluation of videotaped ITAs' actual classroom teaching sessions. The further to the right, the more entangled speaking abilities are with teaching skills, increasing the chances that examinees' speaking abilities are impacted by their teaching skills, and making it difficult for the assessors to separate them out in their evaluations.

The scoring rubrics of these local tests also range from primarily linguistically driven criteria to real-world criteria. For example, the scoring rubric for the UCLA TOP test is most representative of a linguistically driven rubric in which teaching abilities are clearly not scored whereas the rubrics for the other three local ITA tests contain, to varying degrees, teaching abilities or demonstration of an understanding of the American university classroom culture. In the latter case, non-linguistic factors such as personality, rapport with students and concern about students' learning may play important roles. Therefore, the more non-language abilities that the ITA test engaged and the more influence that the non-language components had on the overall evaluation of the ITA test performance, the weaker the relationship was between the TOEFL iBT Speaking Section scores and the ITA test scores.

As it requires a minimal threshold language level for

Table 9. **Summary of the TOEFL iBT Speaking Section Cut Score Recommendations at the Four Institutions**

	Pass	Provisional Pass	Criterion Measure	Cross Validation
UCLA	27	23-24	In-house teaching simulation test	Yes
UNCC	24	Not available[2]	In-house teaching simulation test	Yes
Drexel	23[1]	Not available[3]	ITA course instructor recommendation	No
UF	27-28	23	SPEAK	No

1. For unrestricted teaching assignments, including those requiring large-group instructional contact.

2. At UNCC, students are classified as either pass or fail based on their scores on the local ITA tests.

3. At Drexel, students may be classified into three categories: no instructional contact (NC), restricted assignments (RA), or non-restricted (all) assignments (AA). However, because none of the participants in this study was classified as an NC, it was not possible to establish a cut score for the restricted assignments (RA).

communication strategies to aid communication, the TOEFL iBT Speaking Section, as a test of academic speaking skills, may be an effective measure to screen high level students who are well qualified for teaching and really low level students whose language abilities are below the minimal threshold level. Therefore, it is appropriate to use the TOEFL iBT Speaking Section as an initial pre-arrival screening measure. For borderline students, authentic performance-based tests that require language use in simulated instructional settings may help us to better assess their oral communication skills and their readiness for teaching assignments.

Setting Cut Scores on the TOEFL iBT Speaking Section for ITA Screening

It was found that the TOEFL iBT Speaking Section scores were generally accurate in classifying students into distinct TA assignment groups, the classification accuracy ranging from 71.4% to 96.7% for the model-building samples. For each school, cut scores were recommended in light of the need to minimize the chances of non-passes being classified as passes (Table 9). At UCLA and UNCC, the TOEFL iBT Speaking Section scores were also found to function reasonably well in predicting TA assignments using an independent sample with cut scores determined via the model-building sample.

Conclusion

This study has shown moderately strong relationships between TOEFL iBT Speaking Section scores and local ITA test scores. It has also provided an example of how cut scores can be derived when examinees' performance levels on criterion measures are available. The results have considerable potential value in providing guidance on using the TOEFL iBT Speaking Section scores for ITA screening purposes.

It has to be noted that a recommended cut score for one school being higher than that of another does not necessarily suggest that the former requires stronger speaking skills for their TAs than the latter. The presence of a particular type of student in a sample from a particular institution that did not fit the general prediction model may push up the cut score for that institution as part of the process of minimizing false positives. Thus, an institution needs to think carefully about the characteristics of their ITA population and the kind of language support available before establishing their cut scores.

The TOEFL iBT Speaking Section score recommendations for the four institutions were derived based on the participant samples used in this study. These cut scores need to be closely monitored, validated with new samples in local settings if possible, and modified if necessary. Mechanisms should

be established to rectify cases where ITA assignment classification is not accurate.

Another point worth mentioning is that in this study, the consequences of having potentially unqualified ITAs (false positives) was considered more severe than those of excluding otherwise qualified ITAs (false negatives). Depending on the situation of a particular school, the ITA program may be willing to bear the consequences of having a slightly higher false positive rate to reduce the chances of classifying qualified ITAs as unqualified based on their TOEFL iBT Speaking Section scores. This is certainly a legitimate approach, assuming a mechanism could be established to rectify the situation when unqualified ITAs are put into the classroom, such as setting up a procedure to identify them and then to provide them with the language support they need.

References

Agresti, A. (2002). *Categorial data analysis* (2nd ed.). New York: Wiley.

Educational Testing Service (2002). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Hoekje, B., & K. Linnell. (1994). 'Authenticity' in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly, 28*, 103-125.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Plakans, B. S., & Abraham, R. G. (1990). The testing and evaluation of international teaching assistants. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (pp. 68-81). Washington, D.C.: NAFSA.

Reading Aloud as an Accommodation for a Test of Reading Comprehension

Cara Cahalan Laitusis and Linda L. Cook

Editor's note: *In addition to its own research efforts, ETS is involved in research sponsored by educational institutions and government agencies. ETS researchers are selective about the projects they pursue to ensure that the work is in alignment with ETS's mission to "help advance quality and equity in education by providing fair and valid assessments and related services." The following article describes research supported in part by the U.S. Department of Education to help make large scale assessments of reading proficiency more accessible for students who have disabilities that affect reading. Through such efforts ETS is providing research-based principles and guidelines that help to further the field of education.*

Over the last decade a series of changes to federal legislation have mandated the inclusion of students with disabilities in large scale state accountability assessments (IDEA, 1997, NCLB, 2001, IDEA, 2004). In addition these changes have stipulated that testing accommodations be made, where appropriate, for the inclusion of students with disabilities. Although federal legislation does not distinguish between testing accommodations and testing modifications, most states differentiate between the two and provide a list of each in their guidelines for testing students with disabilities and English language learners. While accommodations are changes to testing procedures or materials that do not alter the construct being assessed or the comparability of test scores (between accommodated and non-accommodated conditions), a testing modification does alter the construct being tested and consequently affects the comparability of test scores. Modifications are sometimes referred to as non-standard administrations or non-allowable accommodations (Thurlow & Wiener, 2000). A recent review of state policy on testing accommodations found that the vast majority of states are in agreement on considering most changes in testing procedures or methods (Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005). For example, nearly all states agree that extra time is an accommodation (not a modification) on state assessments.

States are not in agreement, however, on whether to consider the audio presentation of test content (i.e., read aloud) on reading assessments to be an accommodation or a modification. These differences are largely due to different specifications of reading in each state's reading standards. States that consider read aloud a modification on tests of reading have either (a) determined that reading involves visual or tactile decoding of text or (b) argue that scores are not comparable because the test scores that are obtained with read aloud represent a measure of listening comprehension rather than reading comprehension. On the other hand, states that allow read aloud accommodations on tests of reading or English language arts (ELA) have either (a) defined reading as comprehension of written material that is presented in a visual, tactile, or audio format or (b) allow only portions of the test to be read aloud (e.g., test questions but not passages). Even in states that consider read aloud to be a modification, a significant number of affected students still participate in testing. In California (where read aloud is considered a modification for ELA assessments) more than 5,000 fourth graders (representing nearly 11% of students in special education) took the STAR English Language Arts assessment in 2006 with the test read aloud to the students (Educational Testing Service, 2007). In addition to the division between states in terms of policy there are also frequent changes in testing policy within a state. For example over the last two years at least three states have changed their policy to either allow some form of read aloud (Texas and Massachusetts) or no longer consider read aloud an allowable accommodation (Missouri) on the state reading assessment. States are clearly

Cara Cahalan Laitusis

Linda L. Cook

struggling with the conflict between having reading assessments that are accessible to students with reading-based disabilities and having reading assessments that provide valid measures of the construct of reading.

Starting in 2004 the U.S. Department of Education began funding the National Accessible Reading Assessment Projects (NARAP, see www.narap.info) to research and develop accessible and valid reading assessments for students with reading-based disabilities. The need for accessible reading assessments stemmed from the inconsistency between states in how they defined reading (with or without decoding of print) and a concern that the popular test accommodation of reading a reading test aloud altered the construct being measured and the comparability between test scores from students that had the test read aloud and those that did not. One of the NARAP grants, the Designing Accessible Reading Assessment (DARA) project was awarded to Educational Testing Service in 2004 and has focused a series of research studies on examining the impact of read aloud accommodations on reading comprehension assessments. These research studies have examined the psychometric comparability of test scores that are obtained with and without read aloud accommodations for students with and without reading-based learning disabilities at grades 4 and 8.

It is clear that if a reading assessment is designed to be a direct or indirect measure of decoding (or word recognition), then read-aloud would clearly constitute a test modification. However, it is not clear if audio presentation (read aloud) changes the construct being measured when the construct is defined as comprehension rather than a combination of comprehension and decoding. Phillips (1994) argues that measurement specialists should consider the impact of modifications on the constructs measured and the validity of the resulting test scores. Assuming that an examinee with a disability is incapable of adapting to the standard testing administration, Phillips argues that any changes to testing conditions should be avoided if the change would (a) alter the skill being measured, (b) preclude the comparison of scores between examinees that received accommodations and those that did not, or (c) allow examinees without disabilities to benefit (if they were granted the same accommodation). This last criterion is debatable and several researchers have argued that accommodations should be provided if they offer a "differential" boost to students with disabilities (Elliot & McKevitt, 2000; Fuchs & Fuchs, 1999; Pitoniak & Royer, 2001). More recently, Sireci, Scarpati, and Li (2005) have termed the investigation of this differential performance boost as the "interaction hypothesis." Both the interaction hypothesis and the differential boost argument indicate that an accommodation may still be considered valid if students with disabilities benefit differentially more than students without disabilities when the accommodation is provided. Koenig and Bachman (2004) criticized the differential boost argument as not focusing on the predictive validity of accommodated and non-accommodated test scores and for the potential that ceiling effects can reduce the observed performance gains in the higher performing comparison group. The differential boost design, however, is still considered to be the ideal for examining the validity of test scores obtained with and without accommodations for students with and without disabilities. Based on these theoretical arguments the DARA project researchers designed a series of studies to advance prior research on:

- The validity of test scores (obtained with and without read aloud) within a differential boost framework.
- The comparability of test scores obtained with and without a read aloud accommodation through factor analysis and differential item functioning (DIF) analysis.

Below we will summarize prior research in this area, contributions made by the DARA project, and implications for state testing policy.

Differential Boost

Several studies have used the differential boost framework to examine the impact of audio presentation or read aloud on tests of reading comprehension. Of the five studies reviewed, two found evidence of differential boost from the read aloud accommodation (Crawford & Tindal, 2004; Fletcher, Francis, Boudousquie, Copeland, Young Kalinowski, & Vaughn, 2006) and three did not find any evidence of differential boost (Kosciokek & Ysseldyke, 2000; Meloy, Deville, & Frisbie, 2002; McKevitt & Elliott, 2003). All five studies, however, have one of several limitations; the sample size was too small to detect significant differences, the study did not use a repeated measures design, or the subgroup of students with disabilities was poorly defined. In addition none of the studies examined the validity of test scores taken with and without a read aloud accommodation. While the small sample sizes and repeated measures design are relatively easy to remedy in future research the final limitation "poorly defined disability subgroup" is more difficult to remedy without testing students on their decoding or fluency ability (as Fletcher et al. did).

In response to these limitations the DARA project conducted a large repeated measures designed study of fourth and eighth grade students with and without reading-based learning disabilities. The full sample for this study included 1,181 fourth grade students (527 with reading-based learning disabilities

Table 1. **Design for Reading Comprehension Portion of Differential Boost Study**

	Session 1		Session 2	
Group	Form	Accommodation	Form	Accommodation
1	S	Standard	T	Audio
2	S	Audio	T	Standard
3	T	Standarde	S	Audio
4	T	Audio	S	Standard

and 654 without disabilities) and 847 eighth grade students (376 with reading-based learning disabilities and 471 without disabilities). Each student was randomly assigned to one of four experimental groups displayed in Table 1. The four experimental groups varied the order of administration of test form (S or T) and testing condition (standard or read aloud). Each experimental group was administered one reading comprehension assessment under standard testing conditions and a second reading comprehension assessment with audio presentation (administered via individual CD players). In addition students also completed assessments of reading fluency (both grades) and decoding (fourth grade only). Results from this study indicated that students with reading-based learning disabilities do benefit differentially more from an audio presentation accommodation even after controlling for reading fluency ability and ceiling effects. In addition this study provided some preliminary evidence that test scores obtained from audio presentation accommodations do not predict reading comprehension (based on teachers' ratings of reading comprehension) as well as test scores without audio presentation do for students with reading-based learning disabilities.

Comparability of Scores

While the differential boost research provides valuable information on the validity of test scores obtained with and without read aloud accommodations, additional information is required on the comparability of scores between tests given under these different conditions. Several prior research studies have used data from operational testing programs to either examine the factor structure of tests taken with and without read aloud accommodations or to examine the existence of differential item functioning between groups of test takers who took operational English language arts assessments with or without read aloud accommodations. This section will review the results of prior research and discuss new research studies conducted as part of the DARA project to advance this line of research.

Factor Analysis

Only a small number of studies that have compared the factor structures of assessments given to students with disabilities under accommodated and non-accommodated conditions with scores obtained by students without disabilities are available in the literature. Of this small number of studies, even fewer have examined the factor structure of tests taken with and without read aloud accommodations. Of these studies, one study of a norm-referenced test (Meloy, Deville, & Frisbie, 2002) concluded that the read aloud accommodation appeared to change the construct being measured for most accommodated students relative to the scores of students who were assessed under standard conditions. The authors of the other two studies that examined the factor structure of criterion-referenced state assessments concluded that the results of their studies clearly indicted that, with a small effect size, a one-factor model could be used to describe the data in the accommodated form given to students with a disability and the non-accommodated form given to students without a disability (Huynh & Barton, 2006; Cook, Eignor, Sawaki, Steinberg, & Cline, 2006).

Although these studies provide a consistent finding that audio presentation (read aloud) did not alter that factor structure of criterion-referenced state English language arts assessments, in each study the accommodation and disability samples were confounded. To advance this line of research, the DARA project carried out a series of factor analytical studies, using the same samples and the same reading comprehension tests that were used for the differential boost study described earlier in this report (Cook, 2007). For each grade (4 and 8), the factor structure of the reading comprehension test was compared for; (a) students without a disability who took the test with and without an audio presentation and (b) students with a reading-

based learning disability who took the test with and without the audio presentation. Both exploratory and confirmatory factor analyses were carried out. The results of the studies indicated that the factor structure was very similar for the groups compared. The authors concluded that an audio presentation may not change the construct of reading comprehension as measured by the particular test used in this study.

Differential Item Functioning

In addition to the factor analysis research that has been conducted, several studies have used data from operational testing programs to evaluate the presence of DIF on English/language arts assessments. These studies compare students with disabilities who tested with or without read aloud to students without disabilities who test under standard conditions. The findings from these studies indicate that various levels of differential item functioning exist across the tests studied.

Bolt and Diao (2005) conducted a study to examine whether reading items functioned differently for students with disabilities in grades 4, 5, and 7 receiving a read aloud accommodation than for students without a disability receiving no accommodation. Although several items displayed measurement dissimilarity, there was not a pattern of DIF among the reading subtests uniformly favoring the read aloud group. Similar findings were found in two other studies (Cahalan-Laitusis, Cook, & Aicher, 2004; Lewis, Green, & Miller, 1999).

The Lewis, Green, and Miller (1999) study also compared the degree of DIF on reading and math assessments when the test was read aloud to students with disabilities and when the test was taken under standard conditions by students with and without disabilities. Findings from this study indicated a presence of a larger number of items exhibiting DIF on the reading assessment than on the math assessment, but that the directionality of the DIF did not uniformally favor the read aloud group. Two other studies had similar findings regarding greater presence of DIF when reading tests are read aloud than when math tests are read aloud (relative to reference groups that did not have the test read aloud) but the directionality of DIF was not specifically noted in either study (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Bolt, 2004).

In summary, it appears that English language arts items may function differently for students assessed with accommodations, particularly read-aloud accommodations. However, the differences may at times be counterintuitive, with students tested with the read-aloud accommodation actually performing more poorly than those tested without accommodations. One limitation of this prior research was that disability (both presence and severity) was confounded with accommodation use.

To advance this line of research the DARA project conducted differential item functioning analyses using the same samples and the same reading comprehension test that was used for both the differential boost and factor analysis studies described in this report. (see Cook, Eignor, Sawaki, Steinberg, & Cline, 2007). The DIF studies were carried out using both 4th and 8th grade samples. Two comparisons were carried out for each grade level.

Comparison 1 used students with a reading-based learning disability who took the test under standard conditions as the reference group and students with reading based learning disabilities who took the test with an audio presentation as the focal group.

Comparison 2 used students without a disability who took the test under standard conditions as the reference group and students without a disability who took the test with an audio presentation as the focal group. The results for comparisons 1 and 2 were quite similar. For comparison 1, the results showed only B level DIF and the direction of the DIF was mixed with 9 items favoring the group who took the test without the audio presentation and 7 items favoring the group who took the test with the audio presentation. For comparison 2, 10 of the 48 items in the test displayed DIF with 6 items favoring the students who took the test without the audio presentation and 4 items favoring the students who took the test with the audio presentation.

The researchers concluded that these results supported the results of both the differential boost and factor analysis studies that indicated that an audio presentation may not change the construct measured by this particular reading comprehension test.

Conclusion

The research studies conducted by the Designing Accessible Reading Assessment (DARA) project as well as prior research studies suggest that the audio presentation accommodation does not appear to alter the comparability of test scores. Both the factor analytical studies as well as the differential item functioning studies conducted by DARA indicate that the audio presentation accommodation does not appear to change the construct that the test is measuring. In addition research suggests that students with disabilities, particularly reading-based learning disabilities, do benefit differentially more from read aloud than students without disabilities. However,

preliminary analyses on the validity of test scores indicates that the test scores obtained with read aloud accommodations may not be as predictive of reading comprehension (as it is defined by teachers) as test scores obtained without read aloud for students with reading-based disabilities. Based on these findings we urge states to include a measure of reading fluency when read aloud accommodations are used on tests of reading, particularly if the intent of the assessment is to measure reading fluency indirectly. Future research efforts for the DARA project will explore the feasibility and technical adequacy of developing a multi-stage modified assessment of reading that measures reading fluency, decoding, and comprehension in isolation for students who are unable to demonstrate their proficiency on each of theses skills on an integrated state reading assessment.

References

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis, MN: National Center on Educational Outcomes. Retrieved February 18, 2004, from: http//education.umn.edu/nceo/OnlinePubs/Technical31.htm

Bolt, S. E. (2004, April). *Using DIF analyses to examine several commonly-held beliefs about testing accommodations for students with disabilities.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Bolt, S. E., & Diao, Q. (2005). *Reading-aloud a reading test: Examining reading sub-skill performance.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Cahalan-Laitusis, C., Cook, L., & Aicher, C. (2003, April). *Examining test items for students with disabilities by testing accommodation.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 29, 2006, from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis56.html

Cook, L. (2007, April). *Designing accessible reading assessments for students with disabilities.* Paper presented at the First Annual Oscar K. Buros Lecture. Buros Institute, Lincoln, NE

Cook, L., Pitoniak, M., Cahalan Laitusis, C., & Cline, F. (2007, April). *Examining differential item fuinctioning for a reading test administered with a read-aloud accommodation.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2006, April). *Using factor analysis to investigate the impact of accommodations on the scores of students with sisabilities on English-language arts assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Crawford, L., & Tindal, G. (2004). Effects of a read-aloud modification on a standardized reading test. *Exceptionality, 12*(2), 89-106.

Educational Testing Service (2007). *California Standards Tests (CST) technical report spring 2006 administration* (pp. 158). Princeton, NJ: Author. Retrieved on April 3, 2007, from: www.cde.ca.gov/ta/tg/sr/documents/startechrpt06.pdf

Fletcher, J. M., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., & Vaughn, S. (2006). Effects of accommodations on high stakes testing for students with reading disabilities. *Exceptional Children, 72*(2), 136-150.

Fuchs, L. S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator, 56,* 24-29.

Huynh, H., & Barton, K. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education, 19*(1), 21-39.

Individuals with Disabilities Education Act (IDEA), 20 U.S.C. S 1400 et seq. (1997).

Individuals with Disabilities Education Act (IDEA), 20 U.S.C. S 1400 et seq. (2004).

Kosciolek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Report 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 20, 2006, from: http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm

Koenig, J. A., & Bachman, L. F. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assess.* Washington, DC: National Academies Press.

Lewis, D., Green, D. R., & Miller, L. (1999). *Using differential item functioning analysis to assess the validity of testing accommodated students with disabilities.* Paper presented at the annual CCSSO Large-Scale Assessment Conference, Snowbird, UT.

McKevitt, B. C., & Elliott, S. N. (2003). Effects of perceived consequences of using read-aloud and teacher-recommended testing accommodation on a reading achievement test. *School Psychology Review, 23*(4), 583-600.

Meloy, L. L., Deville, C., & Frisbie, D. A. (2002). The effect of a read-aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education, 23*(4), 248-255.

Meloy, L. L., Deville,C., & Frisbie, D. A. (2002). *The effects of a reading accommodation on standardized test scores of learning disabled and non learning disabled students.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

No Child Left Behind Act of 2001, 20 U.S.C. ž 6301 et seq (2001) (PL 107-110).

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*, 93-120.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research. 71*(1), 53-104.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490.

Thurlow, M., & Wiener, D. (2000). *Non-approved accommodations: Recommendations for use and reporting* (Policy Directions No. 11). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 8, 2005, from: http://education.umn.edu/NCEO/OnlinePubs/Policy11.htm